UCLA College | Physical Sciences
Statistics & Data Science

520 Portola Plaza
8125 Math Sciences Building
Box 951554
Los Angeles, CA 90095

California Fiscal Health Poster

Ruei-Yu Chang (Michelle), Ziwei Guo, Jiarui Song, Shuaiheng Tao, Mandy Vien, Nam Vien
(Department of Statistics & Data Science)

I. Statements & Dataset

High-Level Statement:

We aim to better understand the fiscal risks and challenges faced by California cities to ensure economic efficiency, mitigate potential mismanagement, and prevent waste, fraud, or abuse.

Lower-Level Statements:

1. Cities designated as high-risk often face financial challenges due to inadequate corrective action plans or failure to update fiscal strategies.
2. The interactive dashboard reveals patterns of fiscal distress, including over-reliance on short-term borrowing or declining revenue streams.
3. Audits indicate that proactive monitoring and six-month progress updates are critical to achieving satisfactory fiscal health.
4. Legislative approvals for targeted audits ensure resources focus on cities with the most urgent financial vulnerabilities.
5. Effective corrective action plans are pivotal in removing high-risk designations and restoring economic stability.

Dataset:

The dataset contains filtered financial and fiscal data collected by the California State Auditor's Office to identify high-risk local government agencies and evaluate fiscal challenges faced by California cities.

Figure 1: Auditor State of California Fiscal Health Data

Data Management Statement:

We plan to merge fiscal data from 2016 to 2020 with high-risk designation data from the California State Auditor's Office and anticipate potential data management issues such as missing data, outliers, or any discrepancies in key identifiers used, while ensuring all categorical columns are retained for modeling purposes.

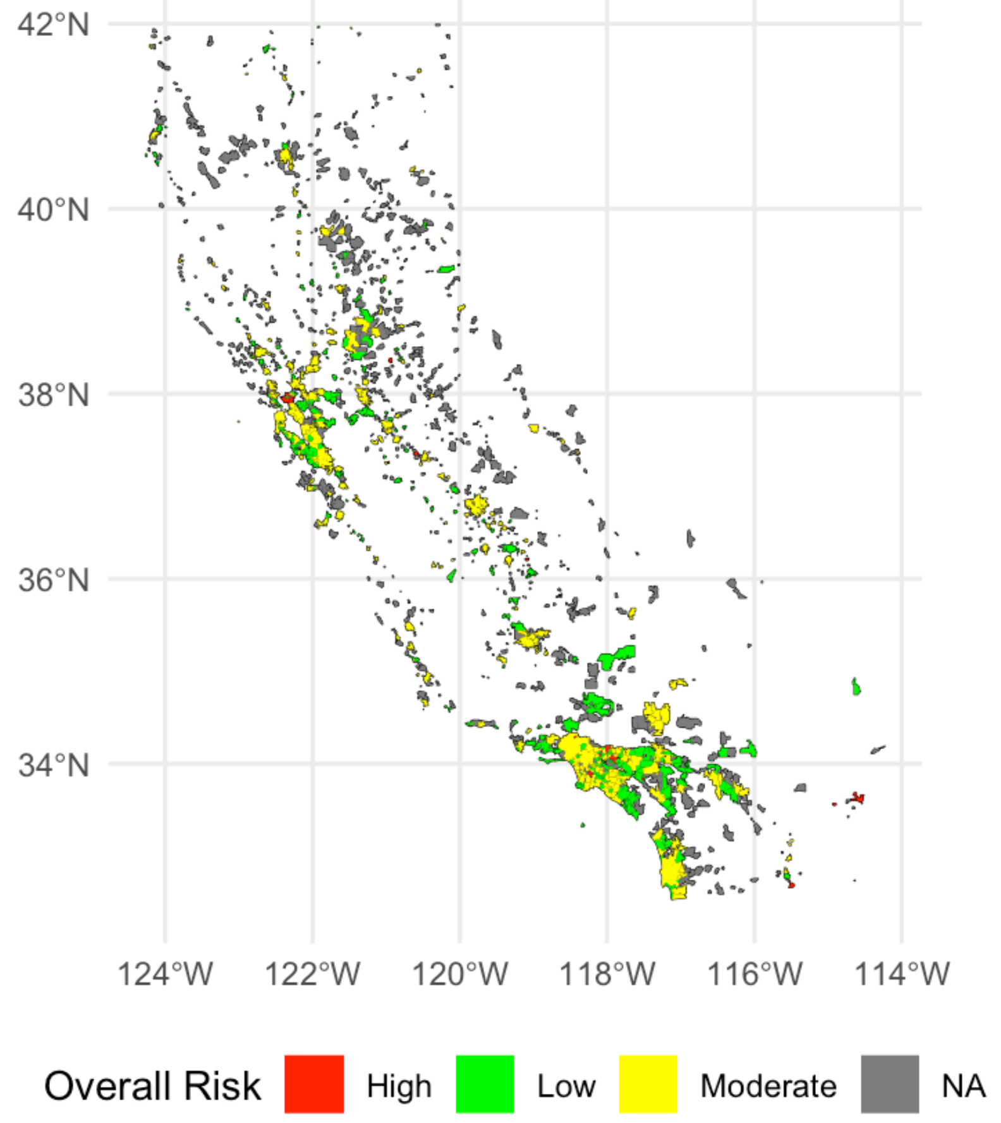

Figure 2: California Cities Overall Risk (2016-2020)

- Revenue trends risk has nearly 0 correlation to overall risk.
- Although there's no missing value in the dataset when we choose to only use the categorical parts. But when we do the map, there are still cities that are NA.

II. Hypotheses for Analysis

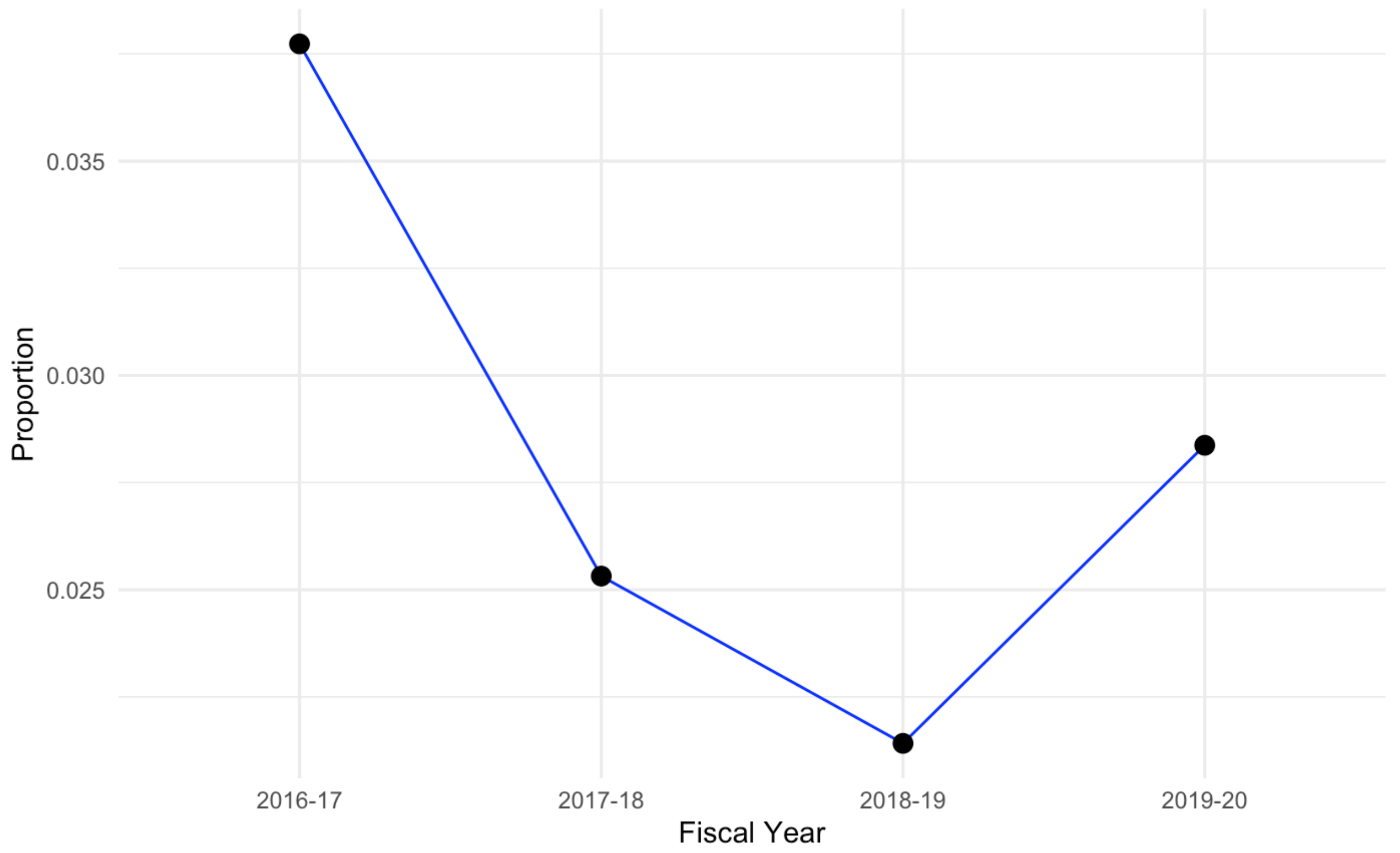

Figure 3: Proportion of Cities with High Overall Risk

1) *Risk Level Trends:*

- (H_0): The proportion of cities categorized as "High Overall Risk" did not change significantly between 2016 and 2020.
- (H_A): The proportion of cities categorized as "High Overall Risk" changed significantly between 2016 and 2020.

The Chi-square goodness-of-fit test resulted in a p-value of 0.4286, which is greater than 0.05. Therefore, we fail to reject the null hypothesis (H_0), indicating no significant evidence that the proportion of cities categorized as 'High Overall Risk' changed across fiscal years.

- (H_0): Year and Overall Risk (all levels) are independent (proportions of all risk categories remain constant across fiscal years).
- (H_A): Year and Overall Risk (all levels) are not independent (proportions vary across fiscal years).

The Chi-Square Test of Independence yielded a p-value of 0.3907, which is greater than 0.05, indicating that we fail to reject the null hypothesis (H_0). This suggests there is no significant evidence of an association between fiscal year and overall risk level.

2) *Pension Risk Association:*

- (H_0): There is no association between "Pension Obligations Risk" and "Overall Risk" levels for cities in the dataset.
- (H_A): There is a significant association between "Pension Obligations Risk" and "Overall Risk" levels for cities in the dataset.

The Chi-square test of independence yielded a p-value $< 2.2e-16$, leading us to reject the null hypothesis (H_0). This indicates that a statistically significant association between Pension_Obligations_Risk and Overall_Risk.

3) *Yearly Variations in Revenue Trends Risk:*

- (H_0): The distribution of "Revenue Trends Risk" categories is consistent across all years in the dataset (2016-2020).
- (H_A): The distribution of "Revenue Trends Risk" categories varies significantly across years in the dataset (2016-2020).

By the Chi-Square Test of Independence, the p-value $< 2.2e-16$ indicates that we reject the null hypothesis (H_0). This suggests that the distribution of Revenue Trends Risk categories varies significantly across fiscal years, with a notable increase in 'Low' risks in 2018–19, while 'Moderate' risks remained consistently high but fluctuated over time.

III. Statistical and Machine Learning Methods

i. Statistical Methods

1) *Risk Level Trends:* Chi-squared test to evaluate changes in proportions over time.
2) *Pension Risk Association:* Chi-squared test to assess the relationship between "Pension Obligations Risk" and "Overall Risk."
3) *Yearly Variations in Revenue Trends Risk:* Chi-squared goodness-of-fit test to analyze changes in categorical distributions across years.

ii. Machine Learning Approach (XGBoost Model)

- Used to predict "Overall Risk" levels based on other risk indicators (e.g., "Pension Obligations Risk," "Revenue Trends Risk").
- Helps identify the most important predictors contributing to high-risk designations, complementing the statistical hypothesis tests.

Why XGBoost over Random Forest?

- Originally, we tried to use numerical values as our features as well. Since there are many missing values, we decided to use XGBoost over the random forest model, as it can automatically handle NAs.
- The overall accuracy of XGBoost is generally higher than that of random forest.
- XGBoost performs faster than the random forest model.

Even though the dataset is highly imbalanced with very few labels for High Overall_Risk, the F1-score for "High" is 0.83, which is a pretty good result and demonstrates the performance of our model.

iii. Integration of Methods

By combining hypothesis testing with the XGBoost model, we aim to:

- Validate trends and associations statistically.
- Use machine learning to explore complex patterns and feature importance in the dataset.

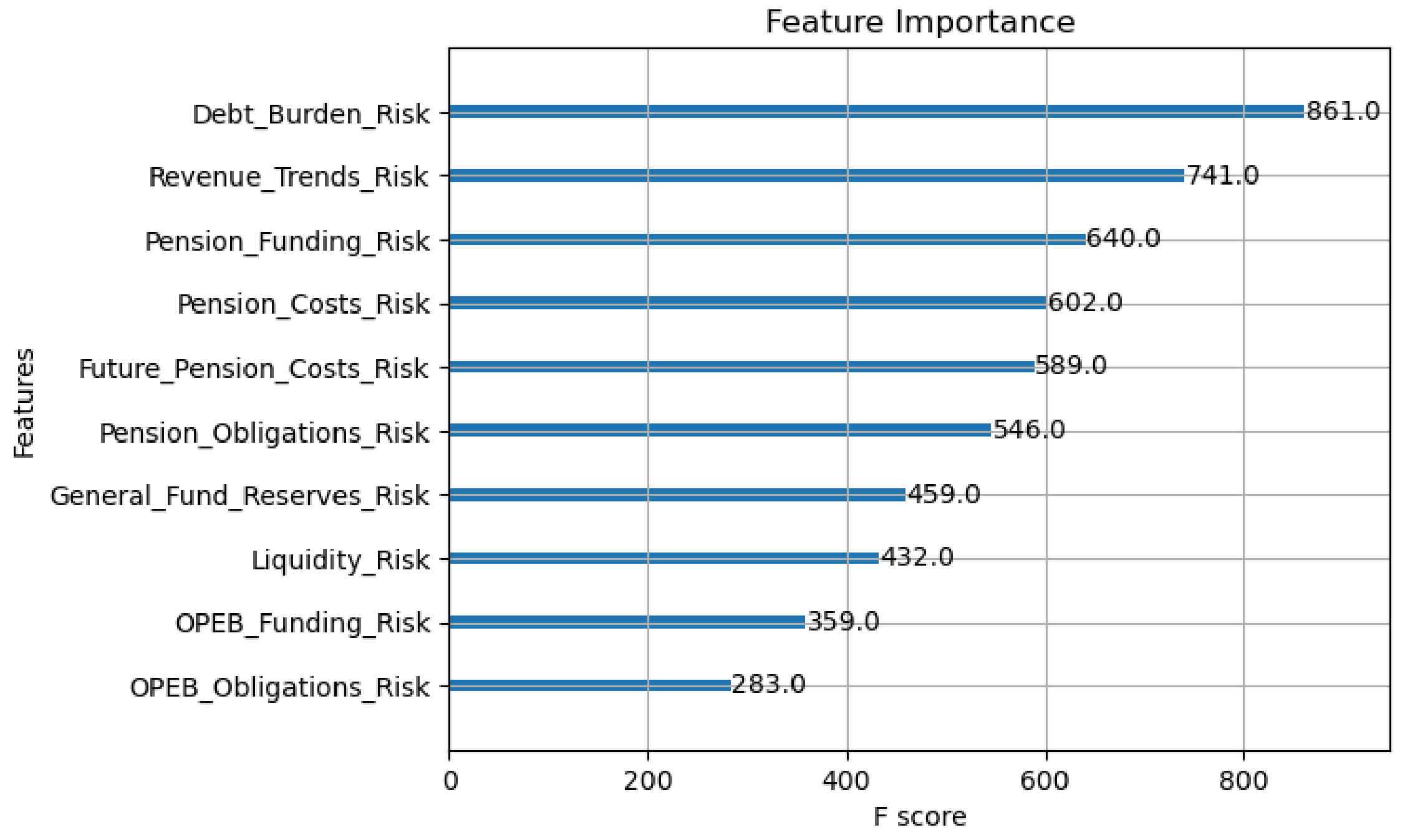

Figure 4: Feature Importance

Findings:

- XGBoost achieved 89.36% accuracy, outperforming Random Forest's 88%.
- The model performed best on Class 1 (moderate risk) with high precision and recall.
- Most misclassifications occurred between Class 2 (low risk) and Class 0 (high risk).
- Key predictors included Debt_Burden_Risk, Liquidity_Risk, and Pension_Funding_Risk.
- Multi-year datasets and categorical encoding supported accurate fiscal risk classification.